Technical Review

# Analysis of Multi-Cloud Data Services for Dell EMC PowerScale Solution for Azure for High Performance File-based Applications

**Date:** April 2021 **Author:** Tony Palmer, Senior Validation Analyst

## Executive Summary

In this Technical Review, ESG audited testing designed to demonstrate how Dell EMC PowerScale for Azure, a part of the Multi-Cloud Data Services for Dell EMC Storage, enabled by Faction—a multi-cloud data services provider—enables demanding file workloads such as large-capacity, high performance computing (HPC) applications including life sciences/bioinformatics and big data analytics, with a focus on performance, utilization, and business value.

***What we found***. ESG validated that Dell EMC PowerScale for Azure gives users with data-intensive workloads both cost-effective storage performance at scale and Microsoft Azure scalable compute performance.

ESG found the solution to be easy to deploy and use, with the end-to-end service, including the Azure ExpressRoute connection and Virtual Machine Scale Sets for automated or manual scaling. HDInsight provided an extremely easy way to deploy Hadoop clusters quickly.

The offering can scale to 90GB/s per Azure virtual network, making it one of the most performant cloud file services we know of in the market today. PowerScale for Azure benefits from network proximity and efficiency with the service delivered from cloud-adjacent data centers.

With PowerScale for Azure, organizations gain advantages of Dell EMC Storage such as high availability for business continuity, data resiliency, and flexible scalability, coupled with the economic benefits of public cloud delivered as an end-to-end managed service. Plus, native replication capabilities allow businesses to move their data from on-premises to workloads in the cloud. If your organization needs to efficiently leverage HPC to solve the most challenging problems in IT today, it would be a smart move to take a serious look at Dell EMC PowerScale for Azure.

### ESG-validated Benefits

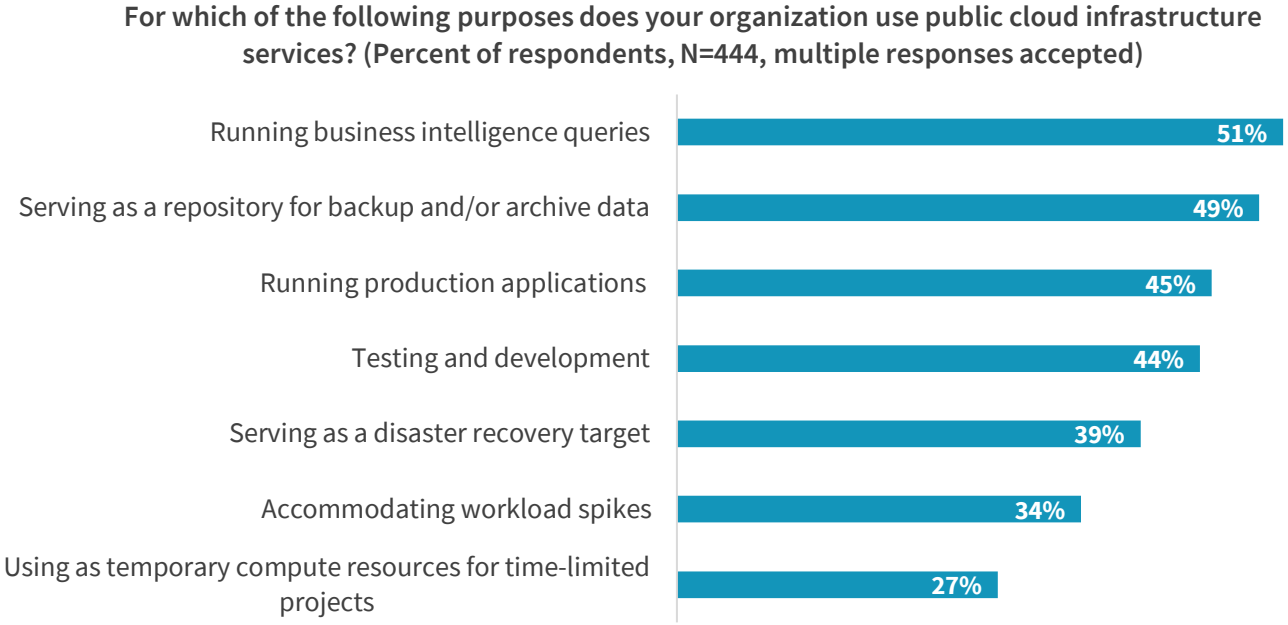
- **Performance and Scale:** 90GB/sec of throughput and tens of petabytes in a single Azure virtual network
- **Cost Efficiency:** Up to 87% lower costs for GPU-based compute than on-premises solutions
- **Ease of Management:** Integrated, managed service with built-in native replication
- **Accessibility:** Multi-protocol access—NFS, CIFS, HDFS, object, and cloud-native service integration
- **All-inclusive Service:** No additional charges and zero egress fees

## Background

Organizations are increasingly moving workloads to the cloud to take advantage of agility, flexibility, and cost-efficiency benefits. According to ESG research, 94% of respondents are currently using public cloud services for infrastructure- or software-as-a-service (IaaS or SaaS). They report that 74% of their current workloads could be targeted to move to the cloud in the next five years. However, only 45% report that they are currently running production applications in public cloud IaaS (see Figure 1).[1]

[1] Source: ESG Research Report, *2020 Technology Spending Intentions Survey*, February 2020.

**Figure 1. What Organizations Use Public Cloud Infrastructure Services For**




*Source: Enterprise Strategy Group*

When we asked respondents what they were using public cloud infrastructure for, the responses they selected apply broadly to both use cases we're examining in this report. Maintaining secondary copies for backup and archive, supplementing production processing for scale-out, test and development, disaster recovery, absorbing workload spikes, and providing temporary capacity for time-limited projects are all key capabilities for enterprises and other organizations using HPC applications.

While file data often accounts for at least half of an organization's on-premises data, unstructured file data is rarely stored in the cloud. This is primarily due to the performance and scale limitations of most solutions available today. In particular, large file workloads such as those used in life sciences, big data analytics, and commercial HPC are not running in the cloud as much as they could be because to date there has been no solution that has offered the levels of performance and scalability these applications demand. And while cloud-resident data analysis tools can be applied to on-premises data, those processes suffer from performance and WAN network bandwidth bottlenecks, which make them both slow and expensive. Other tools, such as file gateways can be deployed, but that increases complexity, adds cost, and may introduce another tier of technical limits incompatible with petabyte-scale HPC workloads.

Extremely compute-intensive processes are one example where customized hardware-accelerated software could provide a competitive difference in cost and/or speed, but deploying dedicated hardware may not make sense as a long-term investment. Using highly optimized graphics processors (GPUs) rather than underpowered, general-purpose CPUs, and optimized hardware-accelerated software can offer a differentiating advantage for an enterprise that can rapidly deploy cloud-based GPUs for a workload.

Dell Technologies has partnered with Microsoft and Faction to offer an Azure-based cloud service that addresses these challenges through cloud-attached enterprise NAS storage that delivers high performance at scale and has a flexible design to optimize costs and keep organizations in control of their data. The Azure solution is a part of the Multi-Cloud Data Services for Dell EMC Storage, which enables organizations to eliminate cloud vendor lock-in by providing a common, consistent data foundation for cloud data that is portable across cloud providers. In this paper we focus on the validation of testing of PowerScale for Azure solution and its benefits.

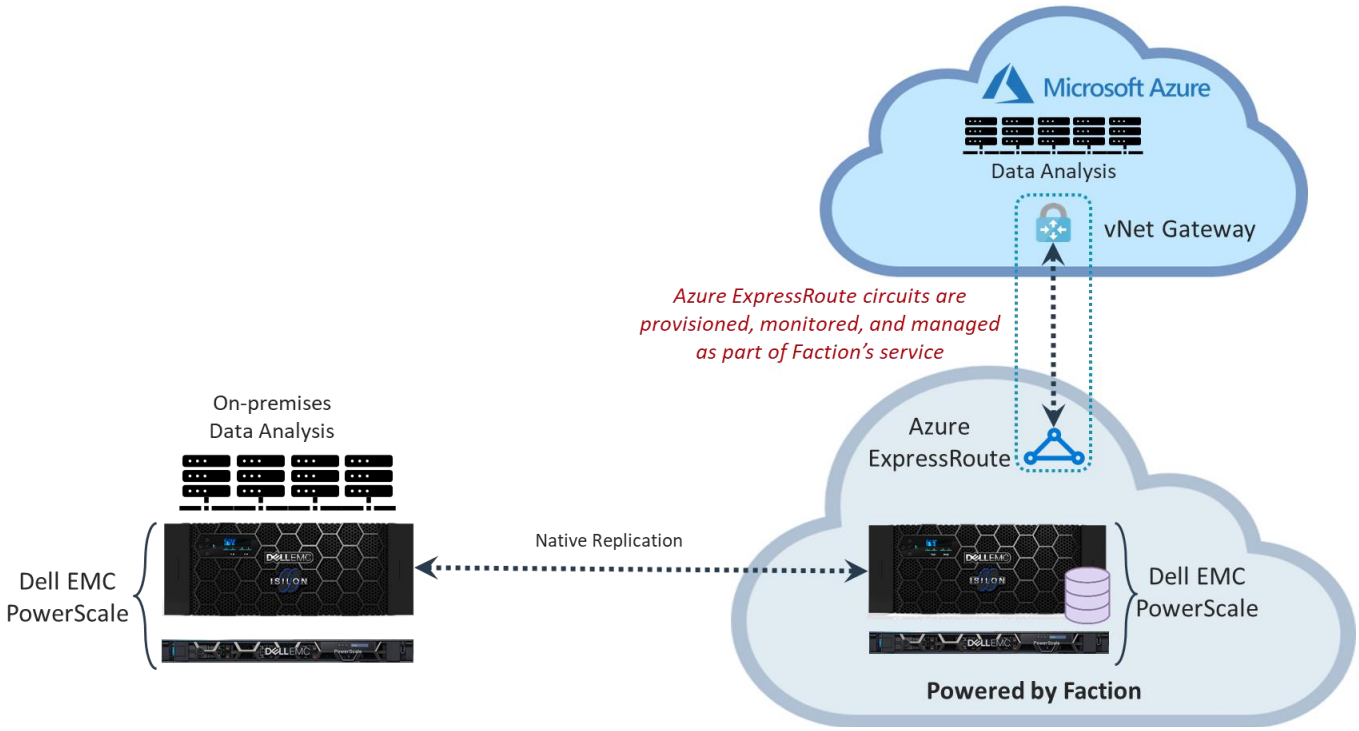
## The Solution: Dell EMC PowerScale for Azure

The PowerScale for Azure solution combines Dell EMC PowerScale with the Microsoft Azure public cloud and Faction's technology to offer enterprise grade compute and storage for operational flexibility. The solution connects PowerScale to Azure via a low-latency, high-bandwidth connection and is designed for data-intensive, high I/O throughput, file-based workloads, especially high-performance computing (HPC) applications such as those related to life sciences, genomics processing, and big data analytics. The service is accessed via standard file protocols, supporting NFS, CIFS, HDFS, as well as object storage over HTTP, and it attaches directly over Azure ExpressRoute with high bandwidth and low latency as to appear effectively LAN-local in the scenarios tested.

Faction, a multi-cloud data services provider, delivers this managed service via a subscription model. The service is available in select regions, but the footprint is expanding. With Faction's Cloud Control Volumes (CCVs) residing on Dell EMC PowerScale storage (Isilon nodes belong to the PowerScale family), organizations can access file-based data via a single multi-petabyte file system in a single namespace. This technology eliminates potential network addressing conflicts as part of the turnkey-managed service. The compatibility of multi-protocol support ensures workflows don't need to be changed—even when the data is used by multiple teams using heterogeneous methods to access it. This is especially important as data sets related to HPC applications tend to grow continually, and the time spent managing data services or coordinating replication or movement of data is no longer available for research or other value-added objectives.

Dell EMC PowerScale nodes are connected at high bandwidth (up to 800Gbps), low latency (as low as .9ms, 1.2ms nominal) to Azure using Azure ExpressRoute Local. Because the connection between Microsoft Azure and Dell EMC PowerScale storage is managed in a cloud-adjacent data center, organizations are not charged cloud egress fees for data written to PowerScale from within Azure, eliminating unexpected usage charges.

**Figure 2. Architecture of the PowerScale for Azure Solution**




*Source: Enterprise Strategy Group*

For computing capabilities, Microsoft Azure offers the choice of dozens of VMs with a wide variety of CPUs and GPUs, some optimized for HPC workloads, memory capacity, and network options. Organizations can choose the right blend of compute instances that will best serve their data processing needs.

## ESG Tested

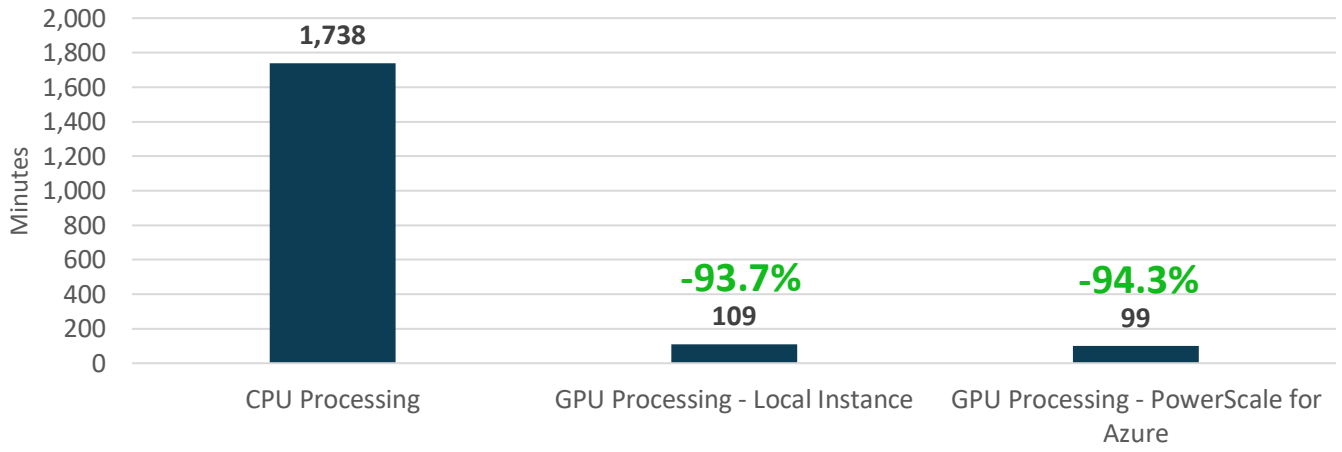
ESG examined PowerScale for Azure across two different use cases requiring HPC for extremely large data sets, including genomic analysis with GPU-optimized software and big data analytics leveraging Hadoop. Azure provides Hadoop services in a turnkey cloud service called Azure HDInsight for which we used the Apache Hadoop Distributed File System (HDFS) data access protocol on PowerScale. Genome analysis is one of the key use cases for life sciences. Genetic information has long-term research value where more data is better for research, but petabyte-scale file repositories are problematic to transfer and expensive to store. Overcoming the financial burdens of massive file repositories, raw processing power, and optimizing the best capabilities of on-premises infrastructure and cloud platforms becomes a top priority.

The raw data generated by a genomic sequencer for the complete genome of a single human is approximately 100 GB before compression. This dictates a requirement for a file system that is massively scalable in both capacity and performance. Genome alignment and sorting, which are both part of the secondary analysis stage, are workloads that demand significant compute, storage, and throughput. Dell Technologies and Azure testing have demonstrated that the performance of PowerScale for Azure scales nearly linearly with increasing I/O demands and a growing number of Azure VMs supporting the genome alignment stage. High bandwidth ExpressRoute Local connections between PowerScale and Azure enable both the compute performance in Azure and the storage performance in PowerScale to process real-world genome analysis.

Data-intensive applications require large volumes of data and can devote most of their processing time to I/O and manipulation of that data. These applications include an array of use cases like predictive financial analysis, business intelligence, healthcare, seismic processing and analysis, geographic information systems, video and media processing, and the emerging autonomous driving and connected cars industry. Most of these applications rely on MapReduce services like Hadoop. MapReduce services can process extremely large data sets with a parallel, distributed algorithm on a cluster of compute nodes. Azure also offers Microsoft Power BI as-a-service to enable self-service business intelligence and meaningful insights with reports and dashboards.

ESG began by measuring the amount of time required to perform a genomic analysis with systems using CPUs and systems using GPUs. "CPU Processing" represents CPU-only genomic analysis, a traditional and commonly used approach that we used as a baseline. We then compared the baseline to two approaches of doing the same analysis but with new GPU technology available in Azure: "GPU Processing–Local Instance," which leverages Azure-based Premium SSD-managed disks, and "GPU Processing–PowerScale for Azure," which leverages the PowerScale for Azure File Scale-Out Elite solution.

**Figure 3. Genomic Analysis—CPU versus GPU-enhanced GATK Software in Azure, Shorter is Better**




*Source: Enterprise Strategy Group*

As seen in Figure 3, leveraging GPU Processing–Local Instance in Azure reduces processing time for our genome data set by 93.7%. Leveraging PowerScale reduces the processing time by 94.3%, or 17.6x faster than CPU processing. PowerScale for Azure delivers features and capabilities that make this solution practical: scalability, multiple access protocols, data protection with snapshots, use of a single copy of the dataset, and no need to copy the data to the cloud.

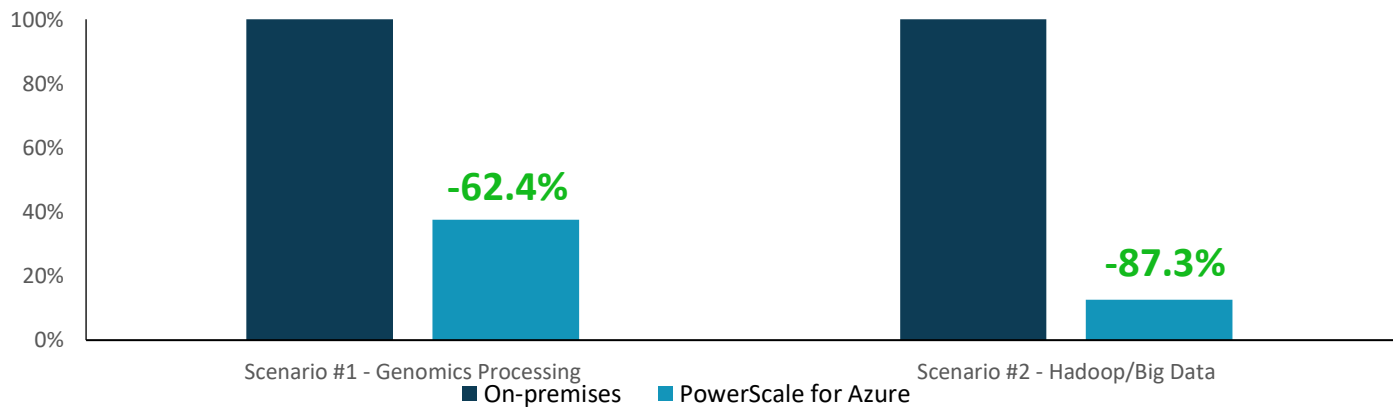
The PowerScale for Azure solution throughput peaked at 140Gb/sec. At 140Gb/sec (17.5GB/sec), this level of throughput is worthy of consideration for many analytics and HPC workloads, with a lot of room to grow for larger environments. For our testing we used the minimum footprint of the PowerScale for Azure File Scale-Out Elite solution, where the maximum footprint is more than 60 times the size of the environment we tested.

Next, we looked at the costs of acquiring and operating compute resources in an on-premises environment versus leveraging PowerScale for Azure. This analysis includes cost of servers, data center space, power, and cooling. We didn't burden on-premises models with any other TCO costs here in terms of design, procurement, or deployment. We looked at two scenarios that could benefit from GPU processing: genomics and big data analytics.

For the on-premises scenarios, we assumed an organization would purchase and maintain 30 systems (120 GPUs) to be able to handle peak processing requirements. We then calculated the costs of running cloud-based workloads based on spot instances in lieu of some or all of the on-premises systems.

In scenario one—a genomics use case—we calculated the cost of acquiring and running 10 systems (40 GPUs) on premises, then bursting to the cloud during peak processing times for just four hours per day using 20 spot instances (80 GPUs). Scenario two—a big data analytics use case—shows much greater savings than does scenario one. This is because scenario two requires an average of eight hours per day of processing with completely variable usage—with no steady state. In this scenario everything is purchased on-demand to satisfy the 100% project-based, highly variable usage, so it's very favorable to running in the cloud (see Figure 4).

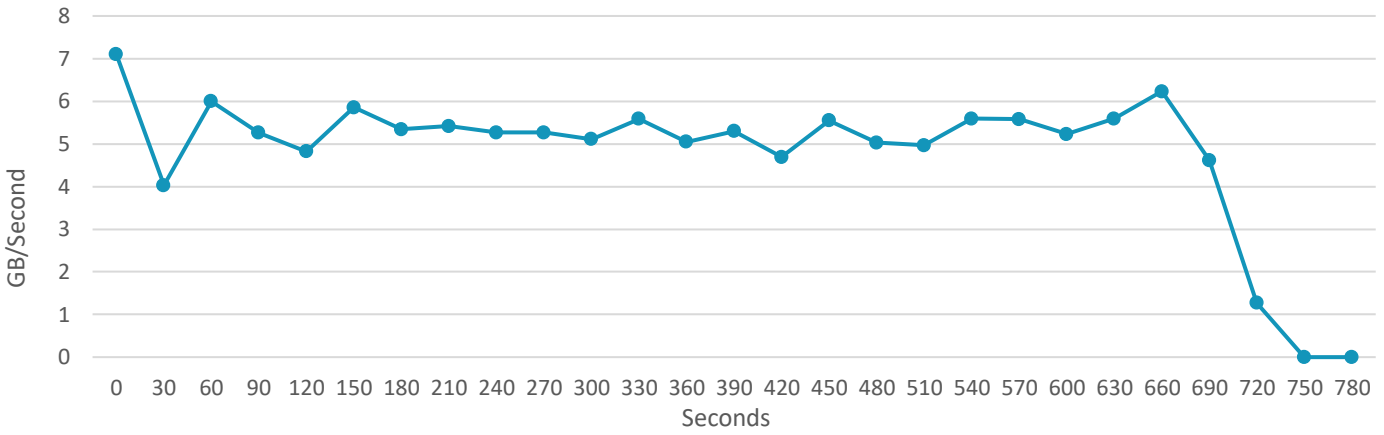
**Figure 4. On-premises versus Dell PowerScale for Azure—Compute Cost Analysis**




*Source: Enterprise Strategy Group*

Finally, ESG audited a series of load tests with Azure HDInsight clusters generating load using the standard Hadoop facilities TeraGen (see Figure 5), TeraSort, and TeraValidate which are a part of the standard Hadoop distribution. The HDInsight service itself is a fully managed Hadoop service in Azure that also supports services such as Apache Spark. The native cloud service in Azure worked seamlessly with the PowerScale for Azure service, integrating with the HDFS protocol offered by the service and functioning identically to other Azure-native services. The test code used generates large data sets, sorts them, and validates the results. ESG focused on the integration between the PowerScale for Azure solution and HDInsight, which is particularly interesting due to fact that although HDInsight is a "cloud native" Azure service, meaning it is fully managed in the cloud and deploys a full stack for Hadoop and analytics, it can still leverage the PowerScale for Azure solution directly using the HDFS protocol.

**Figure 5. Hadoop Ingest Throughput Using TeraGen-28 Worker Nodes**




*Source: Enterprise Strategy Group*

The tests were capped at 28 Standard_E8_v3 nodes representing a total of 224 cores yielding an average throughput of approximately 5.4 GB/sec or 43.2 Gb/sec. Storage throughput scaled linearly with the Hadoop worker node count up to the 28-node configuration with no sign of saturation with 28 worker nodes. ESG validated that HDInsight could be directly configured to utilize the PowerScale for Azure solution as an HDFS target.

## Why This Matters

ESG asked data analytics professionals what aspects of their data pipeline are most frequently responsible for causing delays. Data processing was the most cited answer (26%), with security and governance (19%), preparation (17%), collection (16%), and accessibility (16%) close behind.[2] Computing in the cloud offers the scale needed for processing, while Dell PowerScale for Azure addresses security and compliance—the service is type II soc1/soc2 compliant and HIPAA audited with signed business associate agreements (BAAs), and the multiprotocol single namespace addresses data prep, collection, and access issues.

Even at the minimum footprint, the PowerScale for Azure File Scale-Out Elite performance tier produced impressive throughput. It was observed to deliver over 140Gbp/s of throughput (over 17.5GB/sec). The offering can scale to 90GB/s per Azure virtual network, making it one of the most performant cloud file services we know of in the market today. PowerScale for Azure benefits from network proximity and efficiency with the service delivered from cloud-adjacent data centers. Although the solution is external in nature, extremely low latency was observed, on the order of 0.9 - 1.5ms, depending on the Availability Zone that resources were deployed in.

ESG has validated that PowerScale for Azure scaled linearly when using HDInsight in Azure, scaling to 28 worker nodes for the TeraGen/TeraSort/TeraValidate benchmark where PowerScale for Azure in the Faction DC sustained an average throughput of approximately 5.4 GB/sec or 43.2 Gb/sec.

ESG found the solution to be easy to deploy and use, with the end-to-end service, including the Azure ExpressRoute connection and Virtual Machine Scale Sets for automated or manual scaling. HDInsight provided an extremely easy way to deploy Hadoop clusters quickly. ESG also calculated significant savings—up to 87.3%—in compute costs using cloud spot instances versus purchased hardware. We included minimal operational expenses for space and power. Savings was greatest for the most variable workloads. In the scenarios we tested, Dell EMC PowerScale for Azure provided performance, scalability, and usability benefits across the board thanks to the tight integration with Azure and the ability to leverage features like ExpressRoute, GPU-based Virtual Machines, and HDInsight.

[2] Source: ESG Master Survey Results, *The State of Data Analytics*, August 2019.

## The Bigger Truth

In a research study, ESG surveyed IT and storage professionals responsible for evaluating, purchasing, and managing data storage technology to better understand enterprise storage buying drivers and challenges across both on- and off-premises cloud environments. Nearly half (49%) of respondents stated that their business is reliant on data to some degree, and 93% of storage decision makers identified some level of success in deriving incremental revenue from their data. Three of the top five most-cited applications and workloads that organizations reported that they expected would drive storage spending growth across both on-premises and public cloud infrastructure environments in the next two years fall under the heading of analytics in the form of IoT, data lakes, and AI.[3]

Hybrid cloud strategies involve applying cloud capabilities to on-premises storage infrastructure. When asked to rank several key considerations that go into data storage purchase decisions, more than one-quarter of respondents (26%) identified hybrid cloud as the capability that would have the greatest level of influence.[4] Companies need their IT vendors to not just support hybrid cloud environments but to engineer their solutions to take advantage of hybrid cloud approaches. Only when the entire IT ecosystem supports the hybrid cloud model will organizations fully unlock the power of the public cloud. ESG is seeing evidence of that today as companies move to hybrid cloud environments to meet specific application needs, such as high-performance computing.

Processing petabytes of genomics data benefits from a cloud infrastructure that provides both CPU- and GPU-intensive processing using advanced high-performance computing systems often beyond the business scope of second- and third-tier cloud service providers. This level of performance requires GPU-based resources and service offerings that can optimize results. Taking advantage of readily available resources that are scalable and adaptable is much more flexible—and will not require considerable investment in capital expenditures.

ESG validated multiple benefits of the PowerScale for Azure solution.

- Scale—The ability to massively scale HPC workloads such as genomics or big data analytics, delivering up to 90GB/sec of throughput per file system while supporting tens of petabytes in a single file system.
- Cost efficiency—Up to 87% lower compute costs to acquire and operate than on-premises solutions.
- Ease of use and management—The solution can manage dozens of petabytes in a single file system rather than the hundreds of namespaces required by other solutions. Direct integration with Azure assures ease of consumption, and an end-to-end managed service provides simple and predictable subscription-based pricing. Existing customers of Dell EMC PowerScale benefit from turnkey replication capabilities built into the platform.
- Accessibility—Simultaneous multi-protocol access from thousands of instances.
- Zero egress fees—Avoiding egress charges enables workloads that require a lot of temporary writes to the PowerScale storage nodes to cost-effectively take advantage of Azure's application services.

Dell EMC PowerScale for Azure gives users with data-intensive workloads both cost-effective storage performance at scale and Microsoft Azure scalable compute performance.

With PowerScale for Azure, organizations gain advantages of Dell EMC Storage such as high availability for business continuity, data resiliency, and flexible scalability, coupled with the economic benefits of public cloud delivered as an end-to-end managed service. Plus, native replication capabilities allow businesses to move their data from on premises to workloads in the cloud. If your organization needs to efficiently leverage HPC to solve the most challenging problems in IT today, it would be a smart move to take a serious look at Multi-Cloud Data Services for Dell EMC PowerScale, the solution for Microsoft Azure.

[3] Source: ESG Research Report, *Data Storage Trends in an Increasingly Hybrid Cloud World*, March 2020.
[4] Ibid.